基于多尺度注意力机制的高分辨率网络人体姿态估计 *

李　丽，张荣芬，刘宇红†，陈　娜，张雯雯

(贵州大学 大数据与信息工程学院，贵阳 550025)

摘　要：针对人体姿态估计中面对特征图尺度变化的挑战时，难以预测人体的正确姿势，提出了一种基于多尺度注意力机制的高分辨率网络 MSANet(multiscale-attention net)以提高人体姿态估计的检测精度。引入轻量级的金字塔卷积和注意力特征融合达到更高效的完成多尺度信息的提取；在并行子网的融合中引用自转换器模块进行特征增强，获取全局特征；在输出阶段中将各层的特征使用自适应空间特征融合策略进行融合后作为最后的输出，更充分的获取高层特征的语义信息和底层特征的细粒度特征，以推断不可见点和被遮挡的关键点。在公开数据集 COCO2017 上进行测试，实验结果表明，该方法比基础网络 HRNet 的估计精度提升了 4.2%。

关键词：人体姿态估计；高分辨率网络；多尺度；注意力特征融合；自适应空间特征融合

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2022.03.0109

High resolution network human pose estimation based on multi-scale attention mechanism

Li Li, Zhang Rongfen, Liu Yuhong†, Chen Na, Zhang Wenwen

(*College of Big Data & Information Engineering*, *Guizhou University*, *Guiyang* 550025, *China*)

Abstract: It is difficult to predict the correct human poses when facing the challenge of the scale change of the feature map in the human pose estimation. To solve this problem, proposing a high-resolution network MSANet (Multiscale-Attention Net) based on multi-scale attention mechanism to improve the detection accuracy of human pose estimation. Introduce lightweight pyramid convolution and attention feature fusion to achieve more efficient extraction of multi-scale information; citing the self-transformer module in the fusion of parallel subnets for feature enhancement to obtain global features; in the output stage, The features of each layer are fused using an adaptive spatial feature fusion strategy as the final output, which more fully obtains the semantic information of high-level features and the fine-grained features of low-level features to infer invisible points and occluded key points. Tested on the public dataset COCO2017, the experimental results show that this method improves the estimation accuracy by 4.2% compared with the basic network HRNet.

Key words: human pose estimation; high-resolution network; multi-scale; attention feature fusion; adaptive spatial feature fusion

0　引言

人体姿态估计(human pose estimation)是计算机视觉研究的热点之一，其目的是从给定的图像或视频中去恢复人体关节点的过程，同时也是计算机理解人类动作、行为必不可少的一步。在众多任务中也离不开姿态估计的研究，如视频监控、智能家居和医疗健康等。

近年来，使用深度学习进行人体姿态估计的方法陆续被提出，且达到了远超传统方法[1~3]的表现。2014 年，Toshev 等[4]提出了深度姿态(DeepPose)网络，首次将 2D 人体姿态估计问题由原本的图像处理和模板匹配问题转换为卷积神经网络(CNN)图像特征提取和关键点坐标回归问题。之后，根据单人和多人的研究，分为自下而上(Down-Top)和自上而下(Top-Down)两种方法。

自下而上(Down-Top)[5~7]的方法是先检测出人体关节点，再根据检测出的关节点进行关键点聚类或者图匹配的方法连接成人体骨架。自上而下(Top-Down)[9~12]的方法是首先对图片进行目标检测，找出所有的人，然后将人从原图片中截取后输入到网络中进行关键点检测。2016 年提出的堆叠沙漏网络(SHN)[9]使用多个沙漏网络串行堆叠在一起并对每个沙漏网络进行监督学习，以热图检测的方法进行人体关节点信息的学习，但是这种串行的方法容易丢失部分信息导致检测结果不够准确，并且对有遮挡的图像难以检测关键点；Chen Y 等[10]在 2018 年提出的级联金字塔网络(CPN)则采用自上而下的检测策略，解决了 SHN 造成部分信息丢失的问题。文献[11]提出的 Simple Baselines 相比 SHN 和 CPN 网络结构显得十分简单，同时检测精度较好。2019 年提出的高分辨率网络(HRNet)[12]摈弃了以往的串联方式，采用了并行子网的方式，通过并行多个分辨率的分支，加上不断进行不同分支之间的信息交互，同时达到强语义信息和精准位置信息的目的。然而，尽管 HRNet 在人体姿态估计中，超越了其他所有基于深度学习的方法，但当面临人体占图片比例不同和遮挡严重或重叠时，不能很好地预测人体的正确姿态。为提取多尺度信息，文献[13]提出的金字塔卷积(PYConv)，包含了不同尺度和深度的卷积核，能够增强图像的感受野，同时提取深层和浅层特征，进而确保了多尺度特征的提取，且相比标准卷积，具有较少的参数量和计算复杂度；为解决多尺度特征融合时尺度变化和小目标所带来的问题，文献[14]提出的注意力特征融合(AFF)中的多尺度通道注意力模块解决了在融合不同尺度的特征时出现的问题；文献[15]中采用的自转换器模块(self transformer)通过基于 transformer 的运作方式来提取相同尺度内不同空间之间的非局部交互，获取全局信息，实现特

收稿日期：2022-03-04；**修回日期：**2022-04-22　**基金项目：**贵州省科学技术基金资助项目(黔科合基础-ZK［2021］重点 001)

作者简介：李丽(1996-)，女，贵州毕节人，硕士研究生，主要研究方向为计算机视觉、机器视觉；张荣芬(1977-)，女，贵州贵阳人，教授，博士，主要研究方向为机器视觉、智能硬件及智能算法；刘宇红(1963-)，男(通信作者)，贵州贵阳人，教授，硕士，主要研究方向为计算机视觉智能图像处理、大数据与智能物联(1693623574@qq.com)；陈娜(1995-)，女，贵州遵义人，硕士研究生，主要研究方向为图像语义分割；张雯雯(1997-)，女，贵州铜仁人，硕士研究生，主要研究方向为计算机视觉、机器视觉．

征增强，以解决多分辨率融合的问题；文献[16]提出的自适应空间特征融合(ASFF)，解决了不同层特征之间的冲突问题，在空间上过滤其他层的无用信息,只保留有用信息来进行融合，充分利用了高层特征的语义信息和底层特征的细粒度特征。

通过对以上的研究与学习，针对人体姿态估计中因尺度变化大或遮挡而导致检测结果不够准确的问题，以 HRNet-W32 为姿态估计的基础网络，提出了一种多尺度注意力机制高分辨率网络；针对多尺度特征提取的问题，提出了结合金字塔卷积和注意力特征融合的 Pyaffneck 模块和 Pyaffblock 模块，针对多分辨率融合的问题，融合前引入自转换器模块进行空间特征交互，实现特征增强；并在最后一个阶段中将不同层的特征进行自适应空间特征融合，更加充分的获取不同尺度之间的语义信息和细粒度特征，以此推断被遮挡或重叠的关键点。

1　高分辨率网络

大多数的卷积网络几乎都是从高分辨率到低分率的结构。高分辨率网络(HRNet)则独辟新径,在卷积的过程中将卷积后缩小的网络单独作为一个分支，在整个过程中保持特征图的高分辨率，通过从高分辨率到低分辨率的子网形成多阶段，并将多分辨率子网并行连接的方法。其总体结构分为四个阶段，第一阶段包含一个子网，第二、第三、第四阶段则由多分辨率模块组成，分别包括 2 个、3 个、4 个多分辨率模块，多分辨率模块如图 1 所示。在每一个子网之间通过反复交换信息来进行多分辨率特征的融合，并始终保留先前阶段的分辨率，且 HRNet 最后的输出采用融合后的高分辨率特征表示。

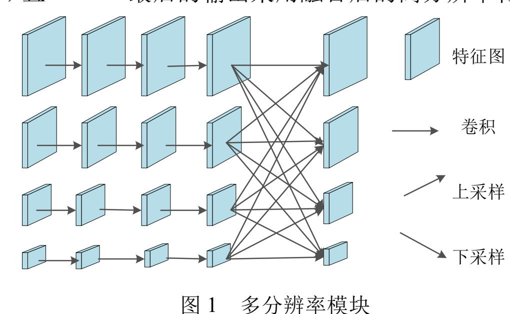


图 1　多分辨率模块

Fig. 1　Multi-resolution module

2　本文方法

本文提出的 MSANet(Multiscale-Attention Net)是基于 HRNet 结构进行改进的，其网络整体结构如图 2 所示。

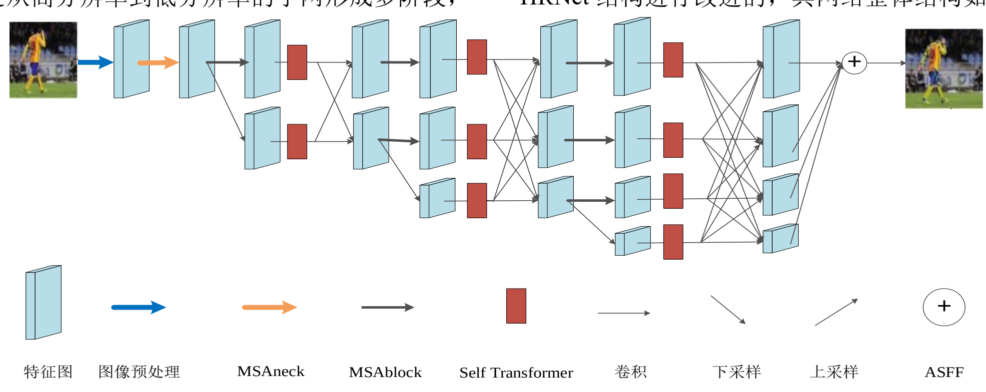


图 2　MSANet 网络结构

Fig. 2　Msanet network structure

MSANet 网络分为 4 个阶段，每个阶段为多分辨率子网的并行连接，且从上到下的子网中，分辨率依次减小 1/2，通道数则依次增加 2 倍。从主干网络开始，由 2 个步长为 3×3 的卷积对图像进行预处理，使分辨率降为原来的 1/4，通道数由原来的 3 变为 64。第一阶段由一个子网构成，使用四个 Pyaffneck 模块来提取特征,并将通道数变为 32。第二、第三、第四阶段则由多分辨率模块构成，分别包含 1,4,3 个多分辨率模块，且每个多分辨率模块通过使用不同分辨率和通道数的 Pyaffblock 模块和自转换器模块(ST)来提取特征。不同于 HRNet，本文将第四阶段输出的四个特征图采用自适应空间特征融合(ASFF)的方法进行融合后作为最后的输出。

本文通过结合金字塔卷积和注意力特征融合构造出 Pyaffneck 模块和 Pyaffblock 模块，将其作为基础模块，有效的提取图像的多尺度特征；然后在融合阶段采用自转换器模块实现跨空间的特征交互，即提取相同尺度内不同空间之间的非局部交互，更有效的提取和融合特征；最后通过上采样操作和自适应空间特征融合将经过反复交换的信息以高分辨率表征的形式输出，实现对人体关键点的检测，从而进一步实现人体姿态估计任务。

2.1　多尺度特征提取

对于人体姿态估计中关键点的多尺度特征的提取，本文将 HRNet 的 bottleneck 模块和 basicblock 模块中的 3×3 卷积替换为金字塔卷积，为克服融合不同尺度的特征时出现的问题，本文使用 AFF 模块进行融合，提出的 pyaffneck 模块和 pyaffblock 模块如图 3 所示。

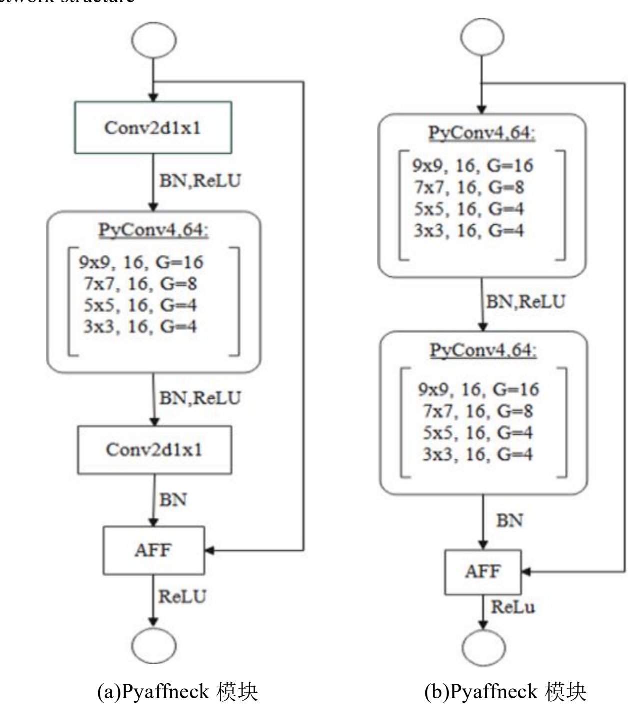


图 3　Pyaffneck 模块和 Pyaffblock 模块

Fig. 3　Pyaffneck module and Pyaffblock module

自深度学习以来，通常使用具有较小内核的卷积神经网络来提取特征，通常为 3×3 卷积，而多尺度特征的提取在于感受野的大小，感受野的大小由卷积核的大小决定，卷积核

越大，感受野越大，看到的图片信息越多，因此获得的特征越好。然而，普通卷积中增加卷积核的大小会导致计算量的增加和计算性能的降低，且普通卷积单一空间大小的单一类型的核，不能提取图像的多尺度特征。HRNet 的 bottleneck 模块和 basicblock 模块中均使用普通卷积来提取特征，使得网络一定程度上不能够准确地出检测小尺度的目标人体及正确的关键点，本文受文献[13]的启发，采用金字塔卷积替换 HRNet 的 bottleneck 模块和 basicblock 模块中的 3×3 普通卷积，以提取图像中的多尺度信息。

如图 4 所示，为尽可能的降低 PyConv 的计算量，使用分组卷积将输入特征分为不同的组，并为每个输入特征组独立应用内核。对于图 4(a)，G=1，此时为标准卷积，每个输出特征都连接到所有的输入特征；图 4(b)，G=2，此时将输入特征映射分为两组，并将每组使用独立的核，使得核的深度减少了 2 倍；图 4(c)则显示当 G=4 时，核的深度减少了 4 倍的情况。因此分组数量越多，连通性和核的深度就会越降低，且减少卷积的参数数量和计算成本。因此与标准卷积相比，PyConv 具有较少的计算量和参数数量，且更为灵活和具有可扩展性。

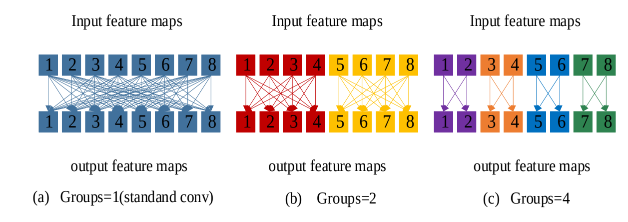


图 4　分组卷积

Fig. 4　Grouped convolution

如图 5，金字塔卷积(Pyraimidal Convolution, PyConv)与标准卷积的区别在于其包含一个核金字塔，其中每一层为不同大小和深度的核，扩大感受野的同时还能使用不同的内核大小来提取图像中多尺度的细节信息。如图 5(b)所示，对输入的特征图 P_i，金字塔卷积 $\{1,2,3,......,n\}$ 的每一层所对应的不同大小内核 $\{K_1^2,K_2^2,K_3^2,......K_n^2\}$，通过分组的方式得到不同深度的核 $\left\{P_i,\frac{P_i}{\left(\frac{K_2^2}{K_1^2}\right)},\frac{P_i}{\left(\frac{K_3^2}{K_1^2}\right)},......,\frac{P_i}{\left(\frac{K_n^2}{K_1^2}\right)}\right\}$，其中所需的参数为

$$\left\{\begin{array}{l} K_n^2\cdot\frac{P_i}{\left(\frac{K_n^2}{K_1^2}\right)}\cdot P_{on}+......+K_3^2\cdot\frac{P_i}{\left(\frac{K_3^2}{K_1^2}\right)}\cdot P_{o3}+ \\ K_2^2\cdot\frac{P_i}{\left(\frac{K_2^2}{K_1^2}\right)}\cdot P_{o2}+K_1^2\cdot P_i\cdot P_{o1} \end{array}\right\} \qquad (1)$$

FLOPs 数为

$$\left\{\begin{array}{l} K_n^2\cdot\frac{P_i}{\left(\frac{K_n^2}{K_1^2}\right)}\cdot P_{on}\cdot(W\times H)+...... \\ +K_3^2\cdot\frac{P_i}{\left(\frac{K_3^2}{K_1^2}\right)}\cdot P_{o3}\cdot(W\times H) \\ +K_2^2\cdot\frac{P_i}{\left(\frac{K_2^2}{K_1^2}\right)}\cdot P_{o2}\cdot(W\times H) \\ +K_1^2\cdot P_i\cdot P_{o1}\cdot(W\times H) \end{array}\right\} \qquad (2)$$

输出特征图 $\{P_{o1},P_{o2},P_{o3},......P_{on}\}$，且 $P_{o1}+P_{o2}+P_{o3}+......+P_{on}=P_o$，即每一层特征图按通道连接得到输出特征图。

金字塔卷积的每一层包含不同尺度和深度的卷积核，不同的卷积核可以有不同的感受野，较小感受野的内核可以关注细节信息来捕捉小目标，增加内核的大小可以捕捉对较大目标更可靠的细节信息，且网络具有可探索性。利用这种卷积方式能在降低计算复杂度和减少参数量的情况下能够同时提取深层和浅层特征，使网络得到了并发性的提高。

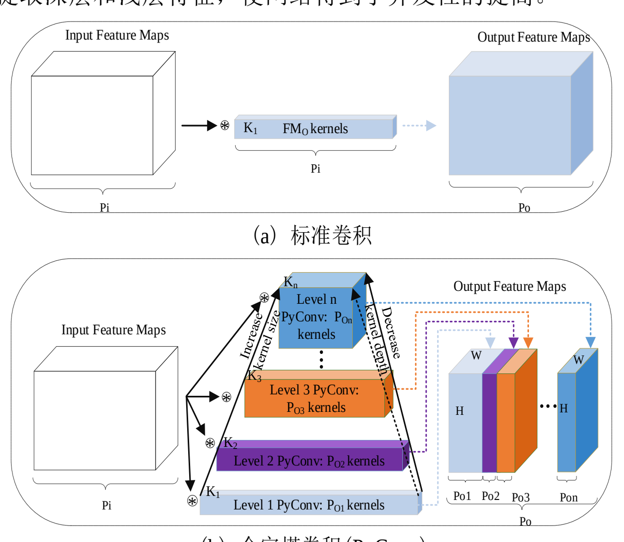


图 5　金字塔卷积

Fig. 5　Pyramid convolution

HRNet 的 bottleneck 模块和 basicblock 模块中均使用直接相加来实现特征融合，这种方式对大物体的检测相对敏感，而对于小物体则较差，因此，为更好的融合语义和尺度不一致的特征，本文将 HRNet 的 bottleneck 模块和 basicblock 模块的相加部分使用 AFF 模块替换，使用 AFF 模块不仅能够提取图像的多尺度特征，相对于直接相加来说，还具有较少的参数量。AFF 模块的结构如图 6 所示，在本文中,将输入的特征信息作为 X,通过卷积操作的输出结果作为 Y，且有特征图 $X,Y\in R^{C\times H\times W}$ 为 AFF 的输入。

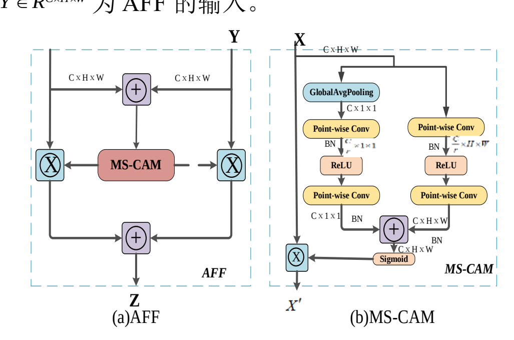


图 6　AFF 结构图

Fig. 6　AFF structure diagram

AFF 的核心模块为多尺度通道注意力模块(MS-CAM)，其结构图如图 6(b)所示，MS-CAM 不是在主干网络中，而是在通道注意力模块中提取局部本地和全局特征的上下文特征。其使用尺度不同的两个分支来提取通道注意力权重，其中一个分支使用全局平均池化(Gloabl Avg Pooling)来提取特征，其计算公式如下：

$$Z=M(X\oplus Y)\otimes X+(1-M(X\oplus Y))\otimes Y \qquad (3)$$

其中，$Z\in R^{C\times H\times W}$ 为融合后的特征，M 为多尺度通道注意模块，$\oplus$ 指相同维度向量的加法运算，$\otimes$ 指向量的乘法运算。

另一个分支为使网络尽可能的较少参数量和计算复杂度，只在注意力模块中将局部上下文添加到全局上下文中，直接使用点向卷积(PWConv)来关注通道的尺度问题，提取局部特征的通道注意力，利用输入特征的每个空间位置的点式通道交互作用，计算公式如下：

$$L(X)=\beta(PWConv_2(\delta(\beta(PWConv_1(z))))) \qquad (4)$$

其中 β 表示 $BatchNom2d(\ )$ 函数，δ 表示激活函数 ReLu，PWConv1 的内核大小为 $\frac{C}{r}\times C\times 1\times 1$,PWConv2 的内核大小为 $C\times\frac{C}{r}\times 1\times 1$。

2.2　多分辨率融合

通常空间上下文信息被动地隐藏在卷积神经网络不断增加的感受野中，或者被 non-local 卷积主动地编码，由于卷积操作是通过不断迭代使用来增大感受野，而这个不断迭代的过程十分低效，不利于最后最优解的求取且只考虑局部区域，忽略了全局其他区域，并不能带来足够的信息。为解决 HRNet 在多分辨率融合阶段中不断的使用上采样和下采样而导致信息丢失等问题，本文采用改进的非局部交互(non-local interaction)自转换器模块(ST)在多分辨率的融合阶段获取全局信息，其输出特征映射与输入特征映射具有相同的尺度，与传统的非局部交互不同的是，使用 Mixture of Softmaxes(MoS)作为归一化函数 ，首先将查询 q，和键 k 分为 N 个部分，然后使用 计算每对图像的相似度得分，基于 MoS 的归一化函数 表达式如下：

$$F_{mos}\left(S_{i,j}^{n}\right)=\sum_{n=1}^{N}\pi_{n}\frac{\exp\left(S_{i,j}^{n}\right)}{\sum_{j}\exp\left(S_{i,j}^{n}\right)} \tag{5}$$

其中，$S_{i,j}^{n}$ 表示第 n 部分的相似度得分，π_n 是第 n 个聚合权重，与 $\mathrm{softmax}\left(w_n^T\bar{k}\right)$ 相等，其中是 w_n 用于归一化的可学习线性向量，$\bar{k}$ 是 k_j 所有位置的算术平均数。

基于 F_{mos}，ST 表示为

Input: q_j,k_j,v_j,N

Similarity: $s_{i,j}^{n}=F_{sim}\left(q_{i,n},k_{j,n}\right)$

Weight:

$$w_{i,j}=F_{mos}\left(s_{i,j}^{n}\right) \tag{6}$$

Output: $X_i=F_{mul}\left(w_{i,j},v_j\right)$

其中，X_i 是 X 中第 i 个转换后的特征位置。

本文采用 ST 模块改进第二、第三、第四阶段的多分辨率融合模块，如图 7 所示，以第三阶段为例，由于 ST 模块能够通过注意力加强距离依赖，扩大感受野，更是直接实现了全局的联系，因此在融合前加入该模块，为后续的信息融合提供更多有用的信息，从而得到更好的融合效果。

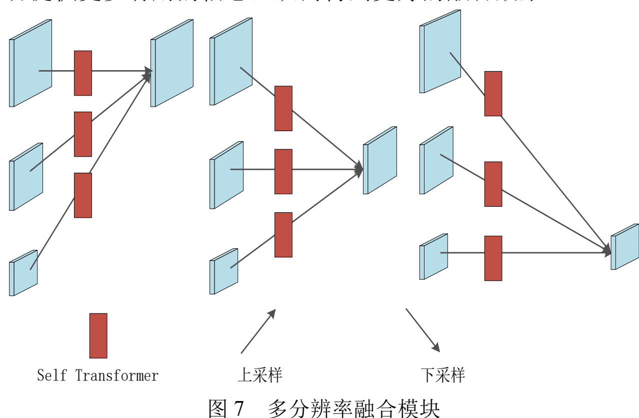


图 7　多分辨率融合模块

Fig. 7　Multi-resolution Fusion Module

2.3　自适应空间特征融合

人体姿态估计中关键点的预测需要较大感受野的具备充分的语义信息的低分变率高层次特征，以推断不可见和被遮挡的关键点，同时，也需要高分辨率的低层次特征进行对某些关键点的进一步细化，以此判断更准的空间位置。为了充分利用高层特征的语义信息和底层特征的细粒度特征，很多网络都会采用金字塔特征表示输出多层特征，然而，不同尺度之间的不一致是基于特征金字塔的单镜头检测器的主要限制，在特征融合时，其他层的很多无用信息也会融合进来。受到文献[16]的启发，为充分利用最后一层 4 种不同大小的特征图，本文采用自适应空间特征融合(ASFF)算法，在最后阶段融合多尺度特征，利用融合后的多尺度信息实现更精确的关键点检测。

ASFF 能够直接学习如何在空间上过滤其他层次的特征，以便只保留有用的信息用于组合，对于某一层次上的特征，首先将其他层次上的特征整合并调整到相同的分辨率，然后训练得到最优融合。本文中 4 种不同大小的特征图分别为原图像大小 1/4、1/8、1/16、1/32，选取 1/4 大小特征图的尺寸和通道数作为融合标准。首先将其他 3 个大小的特征图进行 1×1 卷积，使得通道数转换为与 1/4 大小的通道数一致；其次对于 1/8 大小的特征图，进行 2 倍的上采样，对于 1/16 大小的特征图，进行 4 倍的上采样，对于 1/16 大小的特征图，进行 8 倍的上采样，使得 4 种特征图的大小一致；最后将 4 个特征图 $X_{i,j}^{1}$、$X_{i,j}^{2}$、$X_{i,j}^{3}$、$X_{i,j}^{4}$ 进行自适应空间特征融合，并通过 1×1 卷积后得到最后的输出，使网络始终保持高分辨率表征。

ASFF 的核心思想是通过学习自适应的调整各个尺度特征在融合时的空间权重。本文中调整后的 4 个尺寸、通道数相同的特征图包含了不同的细节信息，ASFF 主要实现根据分配各层的权重参数来融合 4 个特征图，定义 $a_{i,j}$、$b_{i,j}$、$c_{i,j}$、$d_{i,j}$ 为权重参数，则融合策略为

$$a_{i,j}X_{i,j}^{1}+b_{i,j}X_{i,j}^{2}+c_{i,j}X_{i,j}^{3}+d_{i,j}X_{i,j}^{4}=Y_{i,j} \tag{7}$$

其中 $Y_{i,j}$ 为融合后的特征图，$a_{i,j},b_{i,j},c_{i,j},d_{i,j}\in[0,1]$ 且满足：

$$a_{i,j}+b_{i,j}+c_{i,j}+d_{i,j}=1 \tag{8}$$

对于权重参数 $a_{i,j},b_{i,j},c_{i,j}$ 和 $d_{i,j}$ 则是通过将 $X_{i,j}^{1},X_{i,j}^{2},X_{i,j}^{3},X_{i,j}^{4}$ 4 个特征图经过 1×1 卷积得到的，并且参数 $a_{i,j},b_{i,j},c_{i,j}$ 和 $d_{i,j}$ 经过 contact 之后通过 softmax 使得他们的范围在[0,1]内且和为 1，计算公式如下：

$$\begin{aligned}
a_{i,j}&=\frac{e^{X_{i,j}^{1}}}{e^{X_{i,j}^{1}}+e^{X_{i,j}^{2}}+e^{X_{i,j}^{3}}+e^{X_{i,j}^{4}}}\\
b_{i,j}&=\frac{e^{X_{i,j}^{2}}}{e^{X_{i,j}^{1}}+e^{X_{i,j}^{2}}+e^{X_{i,j}^{3}}+e^{X_{i,j}^{4}}}\\
c_{i,j}&=\frac{e^{X_{i,j}^{3}}}{e^{X_{i,j}^{1}}+e^{X_{i,j}^{2}}+e^{X_{i,j}^{3}}+e^{X_{i,j}^{4}}}\\
d_{i,j}&=\frac{e^{X_{i,j}^{4}}}{e^{X_{i,j}^{1}}+e^{X_{i,j}^{2}}+e^{X_{i,j}^{3}}+e^{X_{i,j}^{4}}}
\end{aligned} \tag{9}$$

3　实验与分析

3.1　数据集简述

COCO 数据集是一个大型的、丰富的物体检测、分割和字幕数据集，由 200000 张图片组成，包含 250000 个标注 17 个关键点的人体样本。训练集上包含有 5700 张图像，验证集上含有 5000 张图像，测试集上有 20000 张图像。标注的 17 的关键点分别为：0 鼻子，1 左眼，2 右眼，3 左耳，4 右耳，5 左肩，6 右肩，7 左肘，8 右肘，9 左手腕，10 右手腕，11 左臀，12 右臀，13 左膝，14 右膝，15 左脚踝，16 右脚踝。

3.2　评估标准

本实验在 COCO2017 数据集上对本文的方法进行验证评估，评估方法采用 MS COCO 官方给定的 OKS(Object Keypoint Similarity)进行评估，使用 PCK(Percentage of Correct Keypoints)作为评估指标。

3.3　实验环境与设置

本实验的实验环境为：Python3.8，PyTorch1.7.0，Linux 系统：Ubuntu20.04，显卡：NVIDIA GeForce GTX 3090。并在训练时将数据集中的图像进行预处理，使得大小固定为 256×192，使用 Adam 对网络进行优化，同时将学习率设置为 0.001，训练周期设置为 210，每个 GPU 的批量大小设置为 30。

3.4　实验验证分析

本文将改进的网络 MSANet 在 COCO 2017 数据集上进行实验，并与其他网络在 COCO 2017 数据集上的实验结果进行比较。

如表 1 所示，将本文方法在 COCO 2017 验证集上的实

验结果与其他方法在 COCO 2017 验证集 NLHR 进行对比，实验结果表明本文所提出的网络 MSANet 相对于其他网络在人体姿态估计中取得了最好的效果，与原网络 HRNet-W32 相比，AP^{50} 提高了 5.1%，AP^{75} 提高了 4.1%，AP^{M} 提高了 3.7%，AP^{L} 提高了 3.9%，AR 提高了 2.2%，mAP 提高了 4.2%。可以看出，本文所提出的方法不仅比其他网络的精度高，更是相对于原网络来说提升了关键点检测的精确度。

表 1　COCO VAL 2017 实验结果对比

Tab. 1　Comparison of COCO VAL 2017 experimental results

Methods	Backbone	Input size	Params	GFLOPs	mAP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
CPN	ResNet-50	256×192	27.M	6.20	68.6	—	—	—	—	—
CPN+OHKM	ResNet-50	256×192	27.M	6.20	69.4	—	—	—	—	—
SimpleBaseline	ResNet-50	256×192	34.M	8.90	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline	ResNet-101	256×192	53.M	12.4	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline	ResNet-152	256×192	68.6M	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNet-W32	HRNet-32	256×192	28.5M	7.10	73.4	89.5	80.7	70.2	80.1	78.9
文献[8]	HRNet-32	256×192	30.7M	8.09	74.8	—	—	71.2	81.7	—
文献[21]	HRNet-32	256×192	29.0M	8.20	76.0	93.6	83.7	73.3	83.5	78.9
文献[22]	HRNet-32	256×192	29.1M	7.10	76.7	93.6	84.6	74.0	81.1	81.3
MSANet	HRNet-32	256×192	**28.1M**	**6.90**	**77.6**	**94.6**	**84.8**	**73.9**	**84.0**	**81.1**

表 2 为将本文方法在 COCO 2017 测试集上的实验结果与其他方法在 COCO 2017 测试集上的结果进行对比，其中文献[5]、文献[6]和文献[7]是自下而上的方法，其余的都是自上而下的方法。根据表中对比结果可看出，本文的方法在降低网络复杂度及参数量的前提下精度得到了一定的提升，且对于自上而下和自下而上的方法均具有更高的准确度。

表 2　COCO test-dev2017 实验结果对比

Tab. 2　Comparison of COCO test-dev2017 experimental results

Methods	Backbone	Input size	Params	GFLOPs	mAP	AP^{M}	AP^{L}
文献[5]	-	-	-	-	61.8	57.1	68.2
文献[6]	-	-	-	-	66.7	62.4	72.9
文献[7]	-	-	-	-	70.5	66.6	75.8
CPN	ResNet-50	384×288	-	-	72.1	68.7	77.2
CPN+OHKM	ResNet-50	384×288	-	-	73.0	69.5	78.1
SimpleBaseline	ResNet-152	384×288	68.6M	15.7	73.7	70.3	80.0
HRNet-32	HRNet-48	384×288	28.5M	16.0	74.9	71.3	80.9
文献[8]	HRNet-32	384×288	30.7M	18.2	75.3	71.8	81.3
文献[21]	HRNet-32	384×288	29.5M	15.2	75.2	72.9	82.8
MSANet	HRNet-32	384×288	**27.9M**	**15.0**	**76.1**	**73.2**	**83.6**

本文将 COCO 2017 验证集上大小为 384×288 的图像进行验证，通过计算关键点正确估计的比例 PCK，即计算检测的关键点与其对应的 Groundtruth 间的归一化距离小于设定阈值的比例，并将其与其他网络模型对关键点估计的精确度做对比。表 3 为对比结果，其中 head 表示头部 5 个关节点平均值；shoulder 表示肩部 2 个关节点平均值；elbow 表示肘部 2 个关节点平均值；wrist 表示腕部 2 个关节点平均值；buttocks 表示臀部 2 个关节点平均值；knee 表示膝盖 2 个关节点平均值；ankle 表示脚踝 2 个关节点平均值；average 表示所有关节点平均值。根据表 3 的对比结果可以看出，本文的方法在各个关节点的估计精度上都有一定的提升，且达到了更高的平均估计精度。

3.5　消融实验

本文基于 HRNet 改进的模型具有金字塔卷积、注意力特征融合、自转换器模块和自适应空间特征融合结构进行集成。实验结果证明，本文所提出的方法使得平均精度达到了 4.2% 的提升。为证明模型中各个模块的有效性，本文在 COCO2017 训练集上进行进一步的分析，分析结果如表 4 所示。实验表明，在 mAP 和 Params 指标下，由于金字塔卷积出色的多尺度特征提取性能及其少量的参数量和计算代价，使得网络在参数量降低 4M 的同时性能提升了 1.8%；在金字塔卷积的基础上融入注意力特征融合构建 Pyaffneck 模块和 Pyaffblock 模块作为基础模块，提取不同尺度的细节信息，因为金字塔卷积和注意力特征融合都能多尺度处理特征且都具有更少的参数量，所以在两者的相辅下使得网络在性能上提升了 2.5% 的同时网络参数量降至 23.8M；

表 3 不同方法检测关键点的 PCK 值比较(%)

Tab. 3　Comparison of PCK values of key points detected by different methods

Methods	Backbone	head	shoulder	elbow	wrist	buttocks	knee	ankle	average
SimpleBaselin	ResNet-50	97.0	87.7	86.1	86.6	70.6	82.0	81.8	86.7
SimpleBaselin	ResNet-101	97.1	87.9	87.1	87.7	71.0	83.9	84.3	87.6
SimpleBaselin	ResNet-152	97.5	88.7	87.5	88.0	71.6	84.6	85.1	88.1
HRNet	HRNet-32	97.3	88.7	87.9	88.6	72.2	84.6	85.4	88.3
HRNet	HRNet-48	97.5	88.5	88.4	89.1	71.5	85.4	86.0	88.5
文献[8]	HRNet-32	97.6	89.2	88.5	89.3	73.3	85.2	86.3	88.9
文献[21]	HRNet-32	97.7	89.3	88.6	89.4	73.5	85.4	86.4	89.0
MSANet	HRNet-32	**98.2**	**90.1**	**89.1**	**90.0**	**73.8**	**85.8**	**86.5**	**89.8**

表 4 消融实验结果

Tab. 4　Results of ablation experiments

network					Params	mAP
HRNet	Pyconv	AFF	ST	ASFF		
√	×	×	×	×	28.5M	73.4
√	√	×	×	×	24.5M	75.2
√	√	√	×	×	23.8M	75.9
√	√	√	√	×	26.2M	76.8
√	√	√	√	√	28.1M	77.6

本文在使用构建的 Paffneck 模块和 Pyaffblock 模块提取多尺度特征的基础上使用自转换器模块进行多分辨率的融合，可以看出，由于自换器模块是一种改进的 non-local，其出色的跨空间特征交互能力使得网络在参数量增加 2.4M 的前提下性能上又得到了 0.7%的提升；在以上基础上，添加了自适应空间特征融合模块后，使网络在参数量仅增加 1.9M 的同时性能提升了 0.8%，这是由于融合时在空间上过滤了无用信息，保留有效信息的同时加大了对小尺度目标的识别，并利用语义信息改善了对关键点的检测，且其附加计算成本也相对较小。

这些数据表明模型中各模块的优越性及其出色的性能使本文的方法相对于原网络而言，不仅整体性能提升了 4.2%，参数量也减少了 0.4M。

3.6　可视化实验分析

为表明本文所提出的网络模型 MSANet 在人体姿态估计中因光照、遮挡或重叠、人体占图片尺度较小和图像分辨率较低等影响下，具有一定的鲁棒性和泛化能力及抗干扰能力，本文进行了可视化实验，即将检测出的人体关键点通过可视化将关键点进行连接，并与原网络 HRNet 的可视化结果进行对比，如图 8 所示，包含了多人、遮挡或重叠、分辨率较低以及不同尺度目标的人体姿态估计结果。

其中，(a)和(b)是多人检测，(c)是对人体的背影进行关键点检测，且人体所处环境光线较暗，(d)是对有遮挡的人体背影的关键点检测，(e)是对分辨率较低的人体关键点检测。从图中可以看出，HRNet 网络模型和 MSANet 网络模型在不同的情境下都能够进行人体姿态估计，但当关键点存在遮挡重叠或人体尺度相对较小时，MSANet 网络模型对小尺度的目标更具有敏感性，由(a)、(b)和(e)可以看出，MSANet 网络模型能够检测出 HRNet 网络模型没有识别检测的关键点，从(c)和(d)可以看出，即使在光线较暗、遮挡的条件下，MSANet 网络模型能够对检测出的关键点进行正确的建模，并对建模错误的关键点进行修正，具有较好的泛化能力和抗干扰能力，

更加证明了本文所采用的各模块的优越性。



图 8　人体姿态估计结果

Fig. 8　Human pose estimation results

4　结束语

本文提出了多尺度注意力高分辨率网络，有效提升了人体姿态估计关键点的检测和识别问题。基于高分辨率网络和本文所提出的 Pyaffneck 和 Pyaffblock 两个基础模块的出色的特征提取能力和泛化能力，使得算法学习多尺度特征的表示时得到了有效的提升；在多分辨率融合阶段融入非局部空间交互自转换器模块，使网络改善了多分辨率阶段的特征融合能力；同时对于输出阶段，使用自适应空间特征融合策略可以获取高低层的有效信息，从而更好地推断出遮挡关键点，进而提升了该算法的整体预测准确度。所提出的网络相对于基础网络 HRNet，mAP 综合提升了 4.2%，且在不同环境下，具有一定的鲁棒性和准确度。但所做的工作还有待改进，如何更好地使网络在性能提升的同时降低网络的运算复杂度和参数量或将人体姿态估计运用于动作识别是下一步所需研究的内容。

参考文献：

[1] Andriluka M, Roth S, Schiele B. Pictorial structures revisited: people detection and articulated pose esitimation [C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE, 2009: 1014-1021

[2] Ladicky L, Torr P H S, Zisserman A. Human pose estimation using a joint pixel-wise and part-wise formulation [C]// 2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE, 2013: 3578-3585.

[3] 张显坤, 张荣芬, 刘宇红. 基于二次生成对抗的人体姿态估计 [J]. 激光与光电子学进展, 2020, 57 (20): 335-343. (Zhang Xiankun, Zhang Rongfen, Liu Yuhong. Human pose estimation based on secondary generative confrontation [J]. Advances in Lasers and Optoelectronics, 2020, 57 (20): 335-343.)

[4] Toshev A, Szefedy C. Deeppose: Human Pose Estimation via Deep Neural Networks. [J]. CVPR, 2014: 1653-1660.

[5] Cao Z, Hidalgo G, Simon T, *et al.* OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields [EB/OL]. (2018-12-18) . [2020-04-15]. https://arxiv. org/abs/1812. 08008.

[6] Li J, Wang C, Zhu H, *et al.* CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) . New York: IEEE Press, 2019: 10855-10864.

[7] Cheng B, Xiao B, Wang J, *et al.* Huang and L Zhang, "HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation, "2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2020, pp. 5385-5394, doi: 10. 1109/CVPR42600. 2020. 00543.

[8] 任好盼, 王文明, 危德健, 等. 基于高分辨率网络的人体姿态估计方法 [J]. 图学学报, 2021, 42 (03): 432-438. (Ren Haopan, Wang Wenming, Wei Dejian, *et al.* Human Pose Estimation Method Based on High Resolution Network [J]. Journal of Graphics, 2021, 42 (03): 432-438.)

[9] Newell A. , Yang K. , Deng J. (2016) Stacked Hourglass Networks for Human Pose Estimation. In: Leibe B. , Matas J. , Sebe N. , Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9912. Springer, Cham. Doi: 10. 1007/978-3-319-46484-8_29.

[10] Chen Y, Wang Z, Peng Y, *et al.* Cascaded pyramid network for multi-person pose estimation [C]/ Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) , 2018: 7103-7112. https://arxiv. org/pdf/1711. 07319v2. pdf.

[11] Xiao B, Wu H P, Wei Y C. Simple Baselines for Human Pose Estimation and Tracking [C]// Proceedings of the European Conference on Computer Vision (ECCV) , 2018: 466-481.

[12] Sun K, Xiao B, Liu D, *et al.* Deep High-Resolution Representation Learning for Human Pose Estimation [C]// 2019 IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR) , IEEE, 2019: 5686-5696. DOI: 10. 1109/CVPR. 2019. 00584.

[13] Cosmin Duta, Liu L, Zhu F, *et al.* Pyramidal convolution: Rethinking convolutional neural networks for visualrecognition [J]. arXiv preprint arXiv: 2006. 11538, 2020.

[14] DAI Y, GIESEKE F, OEHMCKE S, *et al.* Attentional Feature Fusion [EB/OL]. [2020-09-10]. https://arxiv. org/pdf/2009. 14082v1. pdf.

[15] DongZhang, HanwangZhang, Jinhui Tang, *et al.* Feature Pyramid Transformer. arXiv: 2007. 09451.

[16] Liu Song-tao, Huang Di, Wang Yun-hong. Learning spatial fusion for single-shot object detection [J]. arXiv: 1911. 09516, 2019.

[17] LIU C X, CHEN L C, SCHROFF F, *et al.* Auto-DeepLab: hierarchical neural architecture search for semantic image segmentation [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) . New York: IEEE Press, 2019: 82-92.

[18] Pavlo Molchanov, Stephen Tyree, Pruning, *et al.* Convolutional neural networks for resource efficient inference [C]// International Conference on Learning Representations (ICLR) , 2017: 1-17.

[19] J Wen, J Chi, C Wu, *et al.* "Human Pose Estimation Based Pre-training Model and Efficient High-Resolution Representation, "2021 40th Chinese Control Conference (CCC) , 2021, pp. 8463-8468, doi: 10. 23919/CCC52363. 2021. 9549849.

[20] 卢健，杨腾飞，赵博，王航英，罗毛欣，周嫣然，李哲．基于深度学习的人体姿态估计方法综述 [J/OL]．激光与光电子学进展：1-27 [2022-01-05]. http://kns. cnki. net/kcms/detail/31. 1690. TN. 20210311. 1622. 003. html. (Lu Jian, Yang Tengfei, Zhao Bo, *et al.* Review of Human Pose Estimation Methods Based on Deep Learning [J/OL]. Advances in Laser sand Optoelectronics: 1-27 [2022-01] -05]. http://kns. cnki. net/kcms/detail/31. 1690. TN. 20210311. 1622. 003. html.)

[21] 罗梦诗，徐杨，叶星鑫．融入双注意力的高分辨率网络人体姿态估计 [J]．计算机工程, 2022, 48 (02): 314-320. DOI: 10. 19678/j. issn. 1000-3428. 0060493. (Luo Mengshi, Xu Yang, Ye Xingxin. High-resolution network human pose estimation with dual attention [J]. Computer Engineering, 2022, 48 (02): 314-320. DOI: 10. 19678/j. issn. 1000-3428. 0060493.)

[22] 孙琪翔，张睿哲，何宁，张聪聪．基于非局部高分辨率网络的人体姿态估计方法 [J/OL]．计算机工程与应用: 1-11 [2022-04-13]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20210420. 1024. 026. html. (Sun Qixiang, Zhang Ruizhe, He Ning, Zhang Congcong. Human pose estimation method based on non-local high-resolution network [J/OL]. Computer Engineering and Applications: 1-11 [2022-04-13]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20210420. 1024. 026. html.)